Toxin genes identified in Nemertean species, Table 3 from Whelan et al (2014) Genome Bio & Evol.(Antarctic Inverts project)

Website: https://www.bco-dmo.org/dataset/671892
Data Type: Cruise Results
Version:
Version Date: 2016-12-27

Project
» Genetic connectivity and biogeographic patterns of Antarctic benthic invertebrates (Antarctic Inverts)

Contributors	Affiliation	Role
Halanych, Kenneth M.	Auburn University	Principal Investigator
Mahon, Andrew	Central Michigan University	Co-Principal Investigator
Copley, Nancy	Woods Hole Oceanographic Institution (WHOI BCO-DMO)	BCO-DMO Data Manager

Table of Contents

Dataset Description

This dataset was published as Table 3 from Whelan et al (2014).

Related Reference: Whelan, N. V., K. M. Kocot, S. R. Santos, and K. M. Halanych. 2014. Nemertean toxin genes revealed through transcriptome sequencing. Genome Biology and Evolution.6 (12):3314-3325. doi: 10.1093/gbe/evu258.

Raw transcriptome data have been submitted to GenBank under accessions SRX731465, SRX731466, and SRX732127.

Methods & Sampling

From Whelan et al (2014):

Specimen Sampling and Sequencing
Malacobdella grossa was collected off Rhode Island by a commercial vessel harvesting the bivalve Arctica islandica as part of the study by Dahlgren et al. (2000). Paranemertes peregrina and T. polymorphus were collected in False Bay, San Juan Island, Washington. RNA extraction, complimentary DNA (cDNA) library preparation, and Illumina sequencing generally followed the methods of Weigert et al. (2014). In brief, we extracted RNA from M. grossa and Para. peregrina using whole animals and from T. polymorphus using the anterior three quarters of the specimen with TRIzol (Invitrogen). RNA was purified with the Qiagen RNeasy kit (Valencia, CA) using on-column DNAse digestion. cDNA libraries were constructed with the SMART cDNA library construction kit (Clontech Laboratories, Mountain View, CA) following the manufacturer's protocol except that the provided 3' oligo was replaced with the Cap-Trsa-CV oligo as per Meyer et al. (2009). Fulllength cDNA was then amplified using the Advantage 2 PCR system (Clontech) with a minimum number of PCR cycles (i.e., 17-

21) and sent to HudsonAlpha Institute for Biotechnology (Huntsville, AL) for library preparation and sequencing on an Illumina HiSeq 2000 using 2x100 bp paired-end (PE) chemistry. Data for five other nemerteans were retrieved from NCBI (table 1; Riesgo et al. 2012).

Transcriptome Assembly

Transcriptome read quality was assessed with the FASTX toolkit (Gordon 2011). Given overall high read quality, sequences were not filtered prior to assembly. All raw data underwent digital normalization using the python script normalize- by-median.py (Brown et al. 2012) with a k-mer size of 20, a desired coverage (i.e., cutoff) of 30, and four hash tables with a lower bound of 2.5x10^9. Trinity version November 2013 (Grabherr et al. 2011; Haas et al. 2013) was utilized for transcriptome assembly of normalized reads for each species with a k-mer size of 25. Raw reads were assembled as PE data except for the two Lineus species and R. lacteus, for which only single-end reads were available.

Assessment of Assembly Quality

A rarefaction curve (Sanders 1968) of assembly statistics was used to evaluate quality and completeness of assemblies from the newly sequenced species because a nemertean reference genome was not available for comparison. Specifically, we removed 10-70% of PE sequences from the end of raw read files to produce datasets with reduced numbers of reads. These subsampled datasets were assembled as above, and N50 and total number of contigs for each assembly was plotted. A plateau on the rarefaction curve would indicate that adding more sequence data would not considerably change the characteristics of the transcriptome assembly. We further measured completeness of each transcriptome assembly with CEGMA 2.4 (Parra et al. 2007), which determines how many of 248 core eukaryotic genes were present in each transcriptome. Core genes annotated by CEGMA are ones that are highly conserved and chosen from the eukaryotic orthologous groups database (Tatusov et al. 2003). An advantage to an approach like CEGMA is that potential differences in which genes are present among assemblies, as a result of expression differences in various tissue types used for cDNA library preparation, is minimized because housekeeping genes should be expressed in virtually all cells.

Transcriptome Annotation

Annotation of each assembled transcriptome was done with the Trinotate annotation suite (http://trinityrnaseq.source forge.net/annotation/Trinotate.html, last accessed April 13, 2014). In brief, TransDecoder (Haas et al. 2013) was first used to predict open reading frames (ORFs) of at least 300 bp. If multiple, overlapping ORFs were present in the same contig, only the longest ORF was retained. In contrast, if multiple but nonoverlapping 300 bp ORFs were identified, all were retained. Thus, two or more ORFs could originate from the same transcript (i.e., ORFs on both forward and reverse strands and/or multiple ORFs on the same strand for long contigs). Untranslated transcripts and translated ORFs were then queried against the Swiss-Prot database (UniProt Consortium 2014) using Basic Local Alignment Search Tool x (BLASTx) and BLASTp, respectively (Altschul et al. 1997), with annotation coming from the best BLAST hit and associated Gene Ontology (GO) terms (Ashburner et al. 2000). Trinotate then used HMMER 3.1 tool hmmscan (Eddy 2001; Finn et al. 2011) and the Pfam-A database (Punta et al. 2014) to annotate protein domains for each predicted protein sequence. Trinotate results were populated into a SQLite database and placed into a tab delimitated file with scripts provided in the Trinotate package and a custom wrapper (available from http://github.com/halocaridina/bioinformatic-scripts, last accessed May 15, 2014). To roughly characterize the protein composition of each nemertean transcriptome, a custom python script (available from http://github.com/ NathanWhelan, last accessed May 15, 2014) was used to place GO terms for each UniProt annotated transcript from Trinotate into Web Gene Ontology Annotation Plotting (WEGO) format, and annotated GO terms were visualized using the WEGO web service (Ye et al. 2006).

Toxin genes were identified based on sequence similarity to previously characterized animal toxins genes under the assumption that sequence similarity is generally indicative of function (Gabaldon and Huynen 2004). Putative toxin genes were initially distinguished if top BLASTx and/or BLASTp hits in the Trinotate output were a previously characterized eukaryotic toxin gene (as defined by Swiss-Prot or presence of a Pfam domain with 'toxin' in the description). Genes of putative viral or bacterial origin were discarded, which eliminated putative toxin genes that may have been from bacterial endosymbionts or from horizontal gene transfer events. Amino acid sequences of remaining transcripts were then manually searched against the NCBI nonredundant GenBank database (nr) and the Pfam protein domain database using the HMMER 3.1 tool phmmer (Eddy 2001; Finn et al. 2011). Annotated sequences initially identified as a toxin gene by Swiss-Prot were further considered a toxin gene if either 1) cross validation produced a significant hit for a toxin domain family in Pfam or 2) if the best annotated hit from the nr database was labeled a toxin gene. In some instances, a toxin as identified by Trinotate and Swiss-Prot did not have a toxin gene as highest hit against the nr database using phmmer, nor did these transcripts possess a toxin protein domain according to Pfam-A; such transcripts were not further considered as they were potential false positives. Toxin genes passing the above filters were then reciprocally queried against the other transcriptomes to identify putative orthologs via BLAST searches. We also queried previously identified nemertean peptide toxins (i.e., Cytotoxin A-III, Neurotoxin B-II, and Neurotoxin B-IV; Kem

1976; Blumenthal et al. 1981) against all transcriptomes with a tBLASTn search.

Gene Tree Reconstruction
Gene trees were inferred for stonefish toxin (SNTX)-like and Plancitoxin-1-like genes because they were found in all nine nemerteans. Putative SNTX and Plancitoxin genes were translated with TransDecoder (Haas et al. 2013) using default settings. Redundant nemertean protein sequences were then removed from the dataset. The SNTX dataset of von Reumont (2014) was added to the nemertean SNTX-like genes. Putative Plancitoxin-1 genes and nontoxic DNase II genes were retrieved from UniProt and GenBank (supplementary figs. S1 and S2, Supplementary Material online). Alignments were done inMAFFT 3 with the E-INS-i algorithm (Katoh and Standley 2013). The appropriate model of protein evolution was selected for each gene using ProtTest 3.4 (LG+ F +Gamma for both genes; Darriba et al. 2011). RAxML 8 (Stamatakis 2014) was used to infer maximum likelihood gene trees, and 1,000 nonparametric bootstrap replicates were performed to assess nodal support. Trees were rooted with nontoxic homologs.

Data Processing Description

BCO-DMO Processing notes:
- added conventional header with dataset name, PI name, version date
- modified parameter names to conform with BCO-DMO naming conventions

[table of contents | back to top]

Data Files

File
Whelan_2014_T3.csv (Comma Separated Values (.csv), 5.83 KB) MD5:0c7540c3619ceeeb0efc3140b272981f Primary data file for dataset ID 671892

[table of contents | back to top]

Parameters

Parameter	Description	Units
species	taxonomic genus and species name	unitless
identifier	specimen? identifier	unitless
Trinotate_Swiss_Prot_Annotation	Trinotate Swiss-Prot Annotation	unitless
Functional_Annotation_and_Pfam_domains	Functional Annotation and Pfam-A database domains	unitless
Putative_Orthologs	Putative Orthologs	unitless

[table of contents | back to top]

Instruments

Dataset-specific Instrument Name	Illumina HiSeq 2000 at HudsonAlpha Institute for Biotechnology (Huntsville, AL)
Generic Instrument Name	Automated DNA Sequencer
Generic Instrument Description	A DNA sequencer is an instrument that determines the order of deoxynucleotides in deoxyribonucleic acid sequences.

Dataset-specific Instrument Name	Advantage 2 PCR system (Clontech)
Generic Instrument Name	Thermal Cycler
Dataset-specific Description	Used to amplify full-length cDNA
Generic Instrument Description	A thermal cycler or "thermocycler" is a general term for a type of laboratory apparatus, commonly used for performing polymerase chain reaction (PCR), that is capable of repeatedly altering and maintaining specific temperatures for defined periods of time. The device has a thermal block with holes where tubes with the PCR reaction mixtures can be inserted. The cycler then raises and lowers the temperature of the block in discrete, pre-programmed steps. They can also be used to facilitate other temperature-sensitive reactions, including restriction enzyme digestion or rapid diagnostics. (adapted from http://serc.carleton.edu/microbelife/research_methods/genomics/pcr.html)

[table of contents | back to top]

Deployments

Halanych_lab_2011-16

Website	https://www.bco-dmo.org/deployment/671488
Platform	Auburn University lab
Start Date	2011-08-01
End Date	2016-07-31
Description	Invertebrate genomics

[table of contents | back to top]

Project Information

Genetic connectivity and biogeographic patterns of Antarctic benthic invertebrates (Antarctic Inverts)

Coverage: Antarctica

Extracted from the NSF award abstract:

The research will explore the genetics, diversity, and biogeography of Antarctic marine benthic invertebrates,

seeking to overturn the widely accepted suggestion that benthic fauna do not constitute a large, panmictic population. The investigators will sample adults and larvae from undersampled regions of West Antarctica that, combined with existing samples, will provide significant coverage of the western hemisphere of the Southern Ocean. The objectives are: 1) To assess the degree of genetic connectivity (or isolation) of benthic invertebrate species in the Western Antarctic using high-resolution genetic markers. 2) To begin exploring planktonic larvae spatial and bathymetric distributions for benthic shelf invertebrates in the Bellinghausen, Amundsen and Ross Seas. 3) To continue to develop a Marine Antarctic Genetic Inventory (MAGI) that relates larval and adult forms via DNA barcoding.

[table of contents | back to top]

Funding

Funding Source	Award
NSF Office of Polar Programs (formerly NSF PLR) (NSF OPP)	PLR-1043745
NSF Office of Polar Programs (formerly NSF PLR) (NSF OPP)	PLR-1043670

[table of contents | back to top]